CONTENT-BASED SEGMENTATION SCHEME FOR DATA COMPRESSION IN STORAGE AND TRANSMISSION INCLUDING HIERARCHICAL SEGMENT REPRESENTATION

ABSTRACT OF THE DISCLOSURE

In a coding system, input data within a system is encoded. The input data might include sequences of symbols that repeat in the input data or occur in other input data encoded in the system. The encoding includes determining a target segment size, determining a window size, identifying a fingerprint within a window of symbols at an offset in the input data, determining whether the offset is to be designated as a cut point and segmenting the input data as indicated by the set of cut points. For each segment so identified, the encoder determines whether the segment is to be a referenced segment or an unreferenced segment, replacing the segment data of each referenced segment with a reference label and storing a reference binding in a persistent segment store for each referenced segment, if needed. Hierarchically, the process can be repeated by grouping references into groups, replacing the grouped references with a group label, storing a binding between the grouped references and group label, if one is not already present, and repeating the process. The number of levels of hierarchy can be fixed in advanced or it can be determined from the content encoded.